Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Disparity Arbiter: Pseudo label mining with student deficiency awareness

Chuang-Wei Liu [a], Mengtan Zhang [b], Nachuan Ma [c], Jin Wu [d], Hanli Wang [a,e,f], Qijun Chen [a], Rui Fan [a,b,g,h,i,j,k,*]

[a] College of Electronic and Information Engineering, Tongji University, 201804, Shanghai, China
[b] Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, 201210, Shanghai, China
[c] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, 214122, Jiangsu, China
[d] School of Artificial Intelligence, University of Science and Technology Beijing, 100083, Beijing, China
[e] School of Computer Science and Technology, Tongji University, 201804, Shanghai, China
[f] Key Laboratory of Embedded System and Service Computing (Ministry of Education), Tongji University, 201804, Shanghai, China
[g] Shanghai Institute of Intelligent Science and Technology, Tongji University, 201210, Shanghai, China
[h] Shanghai Key Laboratory of Intelligent Autonomous Systems, Tongji University, 201210, Shanghai, China
[i] State Key Laboratory of Autonomous Intelligent Unmanned Systems, Tongji University, 201210, Shanghai, China
[j] Frontiers Science Center for Intelligent Autonomous Systems (Ministry of Education), Tongji University, 201210, Shanghai, China
[k] National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University, Xi'an, 710049, Shaanxi, China

ARTICLE INFO

Keywords:
Stereo matching
Disparity confidence
Pseudo disparity labels
Low-quality data

ABSTRACT

Pseudo label-based frameworks have emerged as a dominant paradigm in self-supervised stereo matching. However, existing methods typically rely on a unilateral filtering strategy based solely on teacher disparity confidence, thereby inevitably introducing label noise and redundancy. We attribute this limitation to the dependency of disparity confidence estimators on intermediate stereo matching data structures, which renders teacher and student confidence values numerically inconsistent and incomparable. To address this, we propose Disparity Arbiter (DispArbiter), a framework that explicitly mines pseudo labels with awareness of student deficiencies. DispArbiter comprises a Disparity-Depth Consistency Network that learns from local disparity error ordinal rankings and estimates confidence independently of intermediate stereo matching data, and a conflict arbitration framework that employs a sparse embedding strategy to ensure numerical consistency between confidence scores across disparity maps. DispArbiter enables a direct comparison between teacher and student confidence, facilitating pseudo label generation from even low-quality teacher disparities specifically where the student underperforms. We conduct experiments under two settings: (1) By solving disparity conflicts between supervised methods and enhancing student results with pseudo labels mined from low-quality teacher data, DispArbiter ranks 1st on the KITTI benchmark. (2) Together with a sparsified decisive disparity diffusion method for generating sparse yet reliable teacher disparities, DispArbiter leads to state-of-the-art (SoTA) stereo matching accuracy on the Middlebury benchmark among self-supervised methods. Our source code is available at mias.group/DispArbiter.

1. Introduction

Metric depth perception is fundamental to autonomous driving, robotics, and augmented reality [1–3]. Among existing approaches, stereo matching—a process analogous to human binocular vision—has emerged as a widely adopted solution for its cost-effectiveness [4,5]. Self-supervised stereo matching learning frameworks primarily utilize either photometric consistency or pseudo labels. While photometric methods have been proven to yield erroneous supervision in areas with stereo ambiguities (*e.g.*, repetitive patterns and texture-less regions) [6–8], pseudo label-based approaches have gained prominence by effectively preserving the prior knowledge of pretrained networks.

Pseudo label-based methods typically select predictions from a teacher network as pseudo disparity labels to supervise a student network. While substantial effort has been devoted to elaborating teacher training strategies [9], the subsequent pseudo label generation frameworks remain coarse. In addition to the conventional left-right disparity consistency check (LRDCC) [8], recent works introduce additional confidence modules that compute only teacher disparity confidence and subsequently apply it to a binary filtering process with a threshold hyperparameter to eliminate erroneous pseudo labels [10–12].

* Corresponding author.
E-mail addresses: cwliu@tongji.edu.cn (C.-W. Liu), zmt200110@tongji.edu.cn (M. Zhang), manachuan@163.com (N. Ma), wujin@ustb.edu.cn (J. Wu), hanliwang@tongji.edu.cn (H. Wang), qjchen@tongji.edu.cn (Q. Chen), rui.fan@ieee.org (R. Fan).

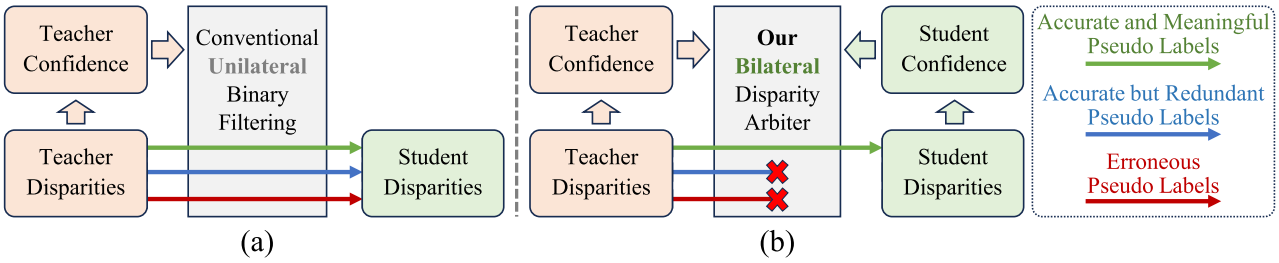

Fig. 1. Comparison of the conventional unilateral binary filtering strategy and our DispArbiter. By jointly evaluating teacher and student confidence, DispArbiter generates pseudo labels exclusively in areas where the student exhibits deficiencies.

However, such a unilateral strategy that focuses exclusively on teacher disparity confidence inevitably introduces label noise, including redundant labels where the student provides similar estimations, and even erroneous labels where the student yields more accurate disparities. Furthermore, given that strictly aligning disparity confidence with the actual error distribution remains intractable [13], teacher disparities exceeding the threshold are not inherently reliable, whereas those excluded may still possess potential training value [14]. Therefore, we are motivated to develop a bilateral strategy that jointly considers both teacher and student confidence, mining informative teacher disparities exclusively in areas where the student exhibits deficiencies, as illustrated in Fig. 1(a).

The restriction of existing pseudo label-based approaches to this unilateral strategy stems from the dependence of their confidence estimation methods on white-box stereo matching [15], where intermediate data structures, such as cost volumes, are accessible [13]. While these internal data provide additional matching cues, the confidence estimation process is tightly coupled with the specific disparity inference process, and the output values are sensitive to the network weights. However, because the teacher and student generally possess distinct weights and even network architectures, their confidence outputs lack numerical consistency, rendering them invalid for direct comparison. Therefore, realizing a bilateral strategy necessitates: (1) performing confidence estimation independently of intermediate stereo data, and (2) ensuring numerical consistency between teacher and student disparity confidence.

To this end, we propose the Disparity Arbiter (DispArbiter), **a framework for explicitly mining pseudo disparity labels with awareness of student deficiencies**. DispArbiter comprises the Disparity-Depth Consistency Network (DDCN), a plug-and-play confidence estimator, and the Conflict Arbitration Framework (CAF), a simple yet effective arbitration mechanism. Specifically, DDCN estimates disparity confidence by evaluating the local pattern similarity between the disparity and relative depth maps. This independence from intermediate stereo data enables deployment on black-box stereo matching [16]. Additionally, a novel Local Error Ranking (LER) loss is proposed to learn the confidence estimation based on the disparity error ordinal rankings from neighboring pixels. Complementarily, CAF employs a sparse embedding strategy that estimates the confidence of conflicting teacher disparities from the student's local patterns to enhance numerical consistency. This enables direct comparison between teacher and student disparity confidence, thereby selectively mining pseudo labels where student disparities exhibit discrepancies, as illustrated in Fig. 1(b).

We conduct experiments under two challenging settings to underscore the generalizability of DispArbiter in terms of student deficiency awareness: (1) By resolving conflicts between supervised stereo matching models with SoTA disparity accuracy, DispArbiter yields even more accurate disparities and **ranks 1st on the KITTI stereo benchmark**. To the best of our knowledge, we present the first attempt to explicitly improve highly accurate disparity maps using less accurate ones. (2) With sparse teacher disparities derived from extremely noisy cost volumes using a sparsified Decisive Disparity Diffusion Stereo (D3Stereo), self-supervised learning with pseudo labels obtained from DispArbiter yields **state-of-the-art (SoTA) accuracy on the Middlebury dataset among all self-supervised methods**. It is noteworthy that although the outputs of DispArbiter are termed pseudo labels, the framework is essentially an accuracy discriminator between any two given disparity maps.

2. Related work

2.1. Disparity confidence estimation

Standard confidence estimation methods predominantly rely on data related to the stereo pipeline, including stereo images, matching cost volumes, and estimated disparities. Peak Ratio Naive [17] and Maximum Margin [18] use cost volumes as the source of confidence information and estimate confidence based on the difference or ratio between cost minima. Confidence Convolutional Neural Network (CNN) [19] introduces CNN into this research domain for the first time and estimates disparity confidence from small patches of the disparity map. Local-Global Confidence (LGC) network [20] leverages a U-Net architecture [21] and estimates disparity confidence from both the image and disparity map. Out-of-The-Box (OTB) [15] imposes a uniqueness constraint by requiring a reliable disparity to establish exclusive correspondence in the reference view.

Recent works exhibit a growing trend toward leveraging intermediate stereo matching data. Mehltretter et al. [22] propose to predict both epistemic and aleatoric disparity confidence based on a Cost Volume Analysis Network (CVA-Net) [23]. SEDNet [13] extracts information from the intermediate multi-resolution disparity maps and introduces a differentiable soft-histogramming technique to approximate the distributions of confidence and disparity errors. Lee et al. [24] propose an alternative to the cost volume that is compatible with black-box stereo matching. Specifically, a sequence of shifted right images derived by disparity plane sweep is fed into a stereo matching network, and the derived sequence of disparity maps is then stacked into a 3D volume for confidence estimation. However, the heavy computational burden imposed by the multiple inference processes renders this strategy computationally prohibitive. Impressively, our previous work DDCV [25] first introduces relative depth maps as a black-box-compatible input modality and subsequently estimates disparity confidence by checking local coherence consistency between disparities and relative depths via several explicit rules. Building upon this concept, we propose DDCN, a learning-based evolution that processes raw disparity and relative depth maps via a siamese network architecture. Moreover, we introduce a novel local error ranking loss specifically designed to learn the local disparity error orders.

2.2. Self-supervised stereo matching

Disparity confidence modules have been increasingly integrated into both photometric consistency-based and pseudo label-based frameworks. In the former category, PASMNet [26] simultaneously regresses confidence and disparity from cost volumes to guide the subsequent

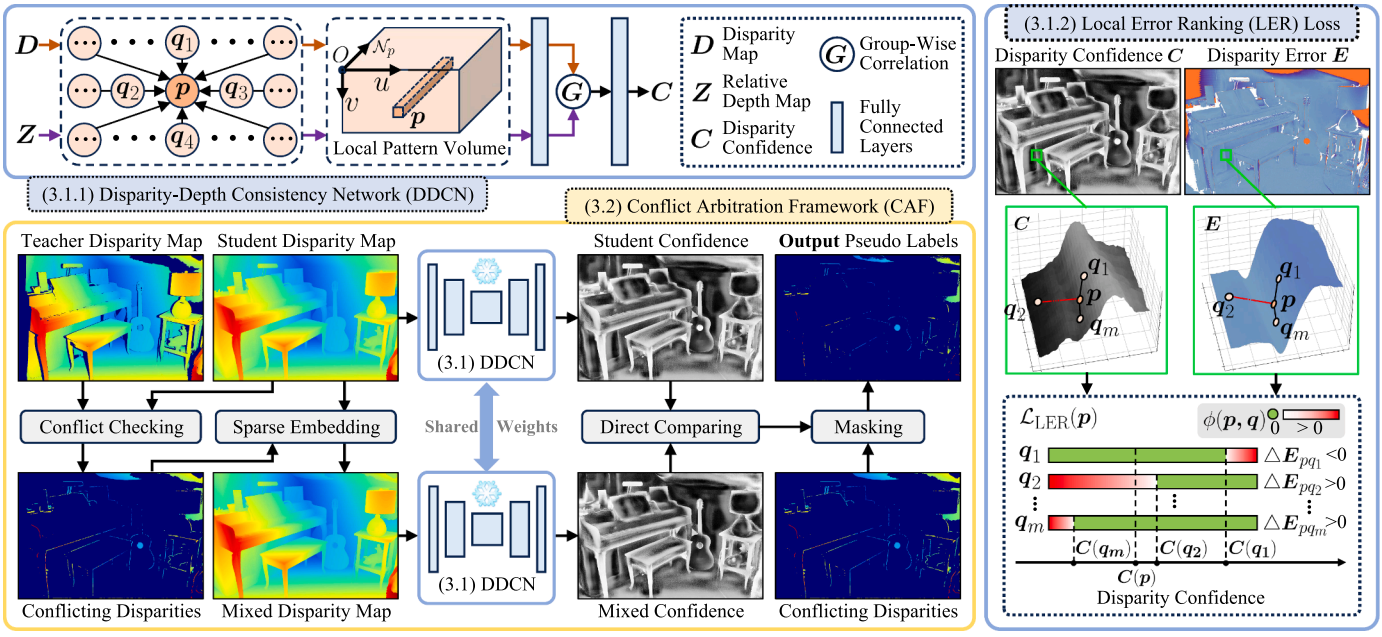

Fig. 2. An illustration of our proposed DispArbiter and LER loss. Correspondences between $\boldsymbol{p}$ and $\boldsymbol{q}_1, \boldsymbol{q}_m$ generate zero loss items $\phi(\boldsymbol{p}, \boldsymbol{q}_1)$ and $\phi(\boldsymbol{p}, \boldsymbol{q}_m)$ in LER loss. DispArbiter, particularly the pre-trained DDCN, is frozen throughout the self-supervised training process of the stereo matching network.

bilinear upsampling refinement. CRD-Fusion [10] combines the Zero-mean Sum of Absolute Difference [27] and Mean Disparity Deviation [28] to generate disparity confidence from both stereo images and raw disparities. The resulting confidence map is then employed to guide an iterative disparity fusion process.

Regarding pseudo label-based frameworks, Semi-Stereo [29] derives confidence from the entropy of disparity distribution within the matching cost volumes, while UCFNet [12] defines pixel-level uncertainty to quantify the multi-modality of the distribution. UnDAF [30] estimates confidence simultaneously with disparity using a single network. Despite these diverse strategies, all these approaches rely on a binary filtering strategy that uses a fixed threshold to filter pseudo labels [14]. Recent methods including DualNet [8] and S^3 [31] leverage data augmentation to differentiate inputs between teacher and student models and thereby enhance the teacher disparity accuracy relative to the student. Afterwards, LRDCC is deployed on the teacher disparities to generate pseudo labels. In this process, LRDCC is also formulated as a confidence estimation process that assigns binarized confidence to teacher disparities. In contrast, our DispArbiter is the first approach to jointly evaluate teacher and student confidence, generating robust pseudo labels that account for student deficiencies.

3. Methodology

As illustrated in Fig. 2, the proposed DispArbiter framework comprises DDCN and CAF. DDCN serves as the cornerstone of DispArbiter for its compatibility with black-box stereo matching. Section 3.1 first details the DDCN architecture and the associated LER loss function. Subsequently, Section 3.2 introduces CAF, which employs a sparse embedding strategy to enforce numerical consistency between confidence scores. Finally, Section 3.3 describes the generation of teacher disparities across two experimental settings and highlights a sparsified D3Stereo algorithm that derives sparse yet accurate teacher disparities from noisy cost volumes. In all experiments, monocular relative depth maps are generated using the Depth Anything 3 model [32].

3.1. Disparity-depth consistency network

3.1.1. DDCN architecture

Recent monocular depth estimation Transformers [33] have been proven to provide robust and accurate local geometric priors. Building upon this, we propose DDCN under the assumption that reliable disparity estimations exhibit local structural consistency with their corresponding relative depths [25]. As illustrated in Fig. 2, DDCN quantifies disparity confidence by evaluating the similarity between local patterns [34] of the disparity map and relative depth map. Specifically, local patterns of the disparity map $\boldsymbol{D}$ and the relative depth map $\boldsymbol{Z}$ are first encoded into local pattern volumes $\boldsymbol{V}_D = \{\triangle \boldsymbol{D}_{pq} | \boldsymbol{q} \in \mathcal{N}_p\}$ and $\boldsymbol{V}_Z = \{\triangle \boldsymbol{Z}_{pq} | \boldsymbol{q} \in \mathcal{N}_p\}$, respectively, where $\triangle \boldsymbol{D}_{pq} = \boldsymbol{D}(\boldsymbol{p}) - \boldsymbol{D}(\boldsymbol{q})$ and $\triangle \boldsymbol{Z}_{pq} = \boldsymbol{Z}(\boldsymbol{p}) - \boldsymbol{Z}(\boldsymbol{q})$ denote the local disparity and relative depth variations, $\mathcal{N}_p = \{\boldsymbol{q}_1, \dots, \boldsymbol{q}_m\}$ is the neighborhood system around pixel $\boldsymbol{p}$ and is constructed using a dilated sampling window. This dilated strategy effectively enlarges the receptive field, capturing broader spatial context at minimal computational cost. Before entering the confidence estimation network, we first align the scales between the disparity and relative depth local patterns with an element-wise scale alignment function $\mathcal{A}(\cdot)$ as follows:

$$\mathcal{A}(\boldsymbol{V}_Z) = \boldsymbol{V}_Z \cdot \frac{\sum_{\boldsymbol{p}\in\boldsymbol{D}} \sum_{\boldsymbol{q}\in\mathcal{N}_p} |\triangle \boldsymbol{D}_{pq}|}{\sum_{\boldsymbol{p}\in\boldsymbol{Z}} \sum_{\boldsymbol{q}\in\mathcal{N}_p} |\triangle \boldsymbol{Z}_{pq}|}, \tag{1}$$

Subsequently, both the local disparity and relative depth pattern volumes undergo processing through an element-wise sign-preserving activation function $\mathcal{F}(\cdot)$ to mitigate excessive variations at depth discontinuities, as formulated below:

$$\mathcal{F}(\boldsymbol{V}) = \mathrm{sgn}(\boldsymbol{V}) \odot \log(1 + |\boldsymbol{V}|), \tag{2}$$

where the $\mathrm{sgn}(\cdot)$ operation preserves the signs of the disparity and relative depth variations, and $\odot$ represents the element-wise Hadamard product between two matrices. The two volumes are then normalized by dividing each by the standard deviation of its absolute values. Afterwards, the processed volumes are passed through fully connected (FC) layers to extract local structural representations. To derive the final confidence scores, these representations are integrated via group-wise cor-

relation (GWC) [35], followed by a regression FC layer. By partitioning channels into compact groups, GWC effectively captures cross-modal similarities between disparity and depth while maintaining the semantic integrity of the features.

3.1.2. Local error ranking loss

In the absence of ground-truth labels for disparity confidence, SED-Net [13] models the ground-truth confidence by aligning the global distributions of confidence and disparity errors via a Kullback-Leibler (KL) divergence loss. However, this global objective is ill-suited for DDCN, as it focuses on local patterns and lacks the global context necessary for matching global distributions. Instead, we propose the LER loss $\mathcal{L}_{\text{LER}}$, which extends the local ranking concept [25] into confidence estimation by learning the ordinal rankings of disparity errors from neighboring pixels. Specifically, $\mathcal{L}_{\text{LER}}$ penalizes neighboring confidence pairs that exhibit ordinal inconsistency with their corresponding disparity errors. By enforcing this ordinal ranking consistency, the LER loss provides supervision compatible with the local patch-based architecture of DDCN. The loss term $\phi(\boldsymbol{p}, \boldsymbol{q})$ between pixel $\boldsymbol{p}$ and one of its neighboring pixels $\boldsymbol{q} \in \mathcal{N}_p$ is defined as follows:

$$\phi(\boldsymbol{p}, \boldsymbol{q}) = |1 - \Theta(\triangle \boldsymbol{E}_{pq} \times \triangle \boldsymbol{C}_{pq})| \times \omega(\triangle \boldsymbol{E}_{pq}) \times \nu(\triangle \boldsymbol{C}_{pq}), \tag{3}$$

where $\triangle \boldsymbol{E}_{pq} = \boldsymbol{E}(\boldsymbol{p}) - \boldsymbol{E}(\boldsymbol{q})$ and $\triangle \boldsymbol{C}_{pq} = \boldsymbol{C}(\boldsymbol{p}) - \boldsymbol{C}(\boldsymbol{q})$ represent the disparity error and confidence variations, respectively. Here, $\Theta(x)$ represents a step function:

$$\Theta(x) = \begin{cases} 1, & x \geqslant 0, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

and the term $|1 - \Theta(\triangle \boldsymbol{E}_{pq} \times \triangle \boldsymbol{C}_{pq})|$ identifies ordinal violations where the confidence variation is inconsistent in sign with the corresponding error variation, as illustrated in Fig. 2. The weighting function is defined as follows:

$$\omega(\triangle \boldsymbol{E}_{pq}) = \frac{|\triangle \boldsymbol{E}_{pq}|}{1 + |\triangle \boldsymbol{E}_{pq}|}, \tag{5}$$

assigns large penalties to pixel correspondences with significant error discrepancies while providing saturated gradients for extreme outliers at depth discontinuities. Similarly, the value term:

$$\nu(\triangle \boldsymbol{C}_{pq}) = \log\left(1 + |\triangle \boldsymbol{C}_{pq}|\right), \tag{6}$$

moderates excessive confidence variations while retaining sensitivity to fine-grained ranking differences. Finally, the LER loss is formulated as follows:

$$\mathcal{L}_{\text{LER}} = \frac{\sum_{p \in \boldsymbol{I}_L} \sum_{\boldsymbol{q} \in \mathcal{N}_p} \phi(\boldsymbol{p}, \boldsymbol{q})}{\sum_{p \in \boldsymbol{I}_L} \sum_{\boldsymbol{q} \in \mathcal{N}_p} |1 - \Theta(\triangle \boldsymbol{E}_{pq} \times \triangle \boldsymbol{C}_{pq})|}, \tag{7}$$

where $\boldsymbol{I}_L$ is the left stereo image. By establishing dense correspondences within local neighborhoods, the LER loss matches the local distributions by ensuring that the predicted disparity confidence aligns precisely with the ordinal ranking of disparity errors, with smaller output values indicating higher confidence.

3.2. Conflict Arbitration Framework

Exploiting the capability of DDCN in estimating disparity confidence from local patterns, our proposed CAF introduces a sparse embedding strategy (SES) to assess the confidence of teacher disparities within student disparity local patterns. By unifying the estimation context, this strategy substantially promotes numerical consistency between the resulting teacher and student confidence scores, rendering them directly comparable, as visualized in Fig. 2. Under this aligned metric, higher confidence scores for each conflicting disparity correspondence reliably signify smaller disparity errors.

Acknowledging that similar teacher and student disparities offer limited learning value and struggle to yield discriminative confidence estimates, CAF exclusively selects conflicting teacher disparities that exhibit considerable deviation from the student predictions as pseudo label candidates. Given teacher disparity map $\boldsymbol{D}^{\text{T}}$ and student disparity map $\boldsymbol{D}^{\text{S}}$, the conflicting disparities $\boldsymbol{D}_C$ are first derived from $\boldsymbol{D}^{\text{T}}$ as follows:

$$\boldsymbol{D}_C = \boldsymbol{D}^{\text{T}} \odot \mathbb{1}\left(|\boldsymbol{D}^{\text{T}} - \boldsymbol{D}^{\text{S}}| > \tau\right), \tag{8}$$

where τ is a manually set threshold to identify conflicting disparities, and $\mathbb{1}(\cdot)$ is an indicator function that retains these conflicting disparities while masking all other positions to zero.

Subsequently, those sparse conflicting disparities are embedded into the student disparity map to generate a mixed disparity map $\boldsymbol{D}^{\text{M}}$. DDCN is then employed to predict the confidence maps $\boldsymbol{C}^{\text{M}}$ and $\boldsymbol{C}^{\text{S}}$ for the mixed and original student disparity maps, respectively. Note that a larger τ reduces the number of conflicting candidates, thereby better preserving the student's local patterns in $\boldsymbol{D}^{\text{M}}$ while risking the omission of potentially useful teacher proposals. Finally, we arbitrate the conflicting disparities by directly comparing $\boldsymbol{C}^{\text{S}}$ and $\boldsymbol{C}^{\text{M}}$, yielding the final pseudo disparity labels $\boldsymbol{D}_P$ as follows:

$$\boldsymbol{D}_P = \boldsymbol{D}_C \odot \mathbb{1}\left(\boldsymbol{C}^{\text{M}} - \boldsymbol{C}^{\text{S}} > \gamma \cdot \beta\right), \tag{9}$$

where $\beta = \text{std}(\boldsymbol{C}^{\text{S}})$ serves as an adaptive scaling factor derived from the confidence distribution, and γ is a hyperparameter that balances the quantity and accuracy of the retained pseudo labels.

3.3. Pseudo label mining

The elimination of grossly inaccurate conflicting teacher disparities prior to their input into DispArbiter further helps prevent distortion of the student's local patterns.

3.3.1. Supervised stereo matching conflict arbitration

When arbitrating disparity conflicts between two supervised stereo matching methods, we designate the less accurate method as the teacher and the more accurate one as the student. Prior to DispArbiter, LRDCC is employed to prune gross outliers from the teacher disparities. The pseudo labels $\boldsymbol{D}_P$ mined from the low-quality teacher results are then substituted into the student disparity $\boldsymbol{D}^{\text{S}}$ to yield an enhanced map $\boldsymbol{D}^{\text{E}}$ as the output of DispArbiter.

3.3.2. Self-supervised stereo matching

In the self-supervised learning of this work, we also propose a sparsified D3Stereo method designed to mine sparse yet reliable teacher disparities. Given explicit matching cost volumes, the original D3Stereo iteratively propagates disparity search ranges from several decisive disparities to their neighbors. For each pixel, D3Stereo retrieves the disparity corresponding to the lowest local cost minimum within the assigned search range. These new estimates are subsequently used to propagate disparity search ranges in the next iteration until no further decisive disparities are produced. Although effective for densification, this unconstrained diffusion process accumulates stereo matching errors, particularly at depth discontinuities.

Our sparsified version introduces two additional constraints to improve accuracy. First, with depth discontinuities identified using an adaptive threshold t_Z defined as follows:

$$t_Z = \frac{\sum_{p \in \boldsymbol{I}_L} \sum_{\boldsymbol{q} \in \mathcal{N}_p} |\triangle \boldsymbol{Z}_{pq}|}{\sum_{p \in \boldsymbol{I}_L} \sum_{\boldsymbol{q} \in \mathcal{N}_p} |\triangle \boldsymbol{D}^{\text{S}}_{pq}|}, \tag{10}$$

disparity diffusion is prevented between correspondences whose depth variations exceed t_Z. Second, we introduce a sparsification strategy, termed Extremum Gating, to prevent the generation and accumulation of erroneous decisive disparities. Specifically, we keep track of the minimum matching cost found thus far for each pixel. This strategy gates the disparity estimation process by accepting a new disparity as a decisive disparity only if its corresponding matching cost is a local minimum and falls below the recorded minimum. Moreover, while the original D3Stereo relies on nearest-neighbor search on low-resolution feature maps for decisive disparity initialization, this work directly initializes

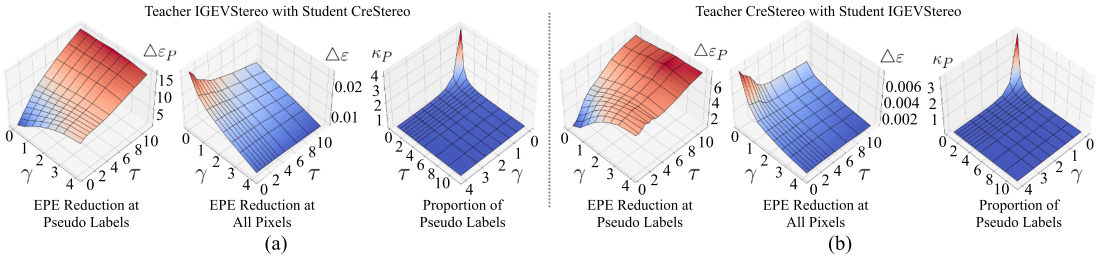

Fig. 3. Experimental results of hyperparameter selection in DispArbiter in terms of the conflicting disparity threshold τ and the ratio hyperparameter γ in the CAF.

Table 1
Ablation studies of the local pattern volume alignment function, activation function, and loss functions. Input disparity maps are generated by our Self-DispArbiter.

$\mathcal{A}(\cdot)$	$\mathcal{F}(\cdot)$	$\mathcal{L}_{\text{LER}}$ **(ours)**	$\mathcal{L}_{\text{GD}}$ [13]	AUC
✔	✔	✔		**0.90**
✔		✔		1.12
	✔	✔		0.93
✔	✔		✔	0.95
✔			✔	1.16
	✔		✔	1.02

them from student disparities via LRDCC. Finally, we feed the disparities from our sparsified D3Stereo and student network into DispArbiter as teacher and student disparities, respectively. The resulting pseudo disparity labels are then used to supervise the self-supervised training of the student network.

4. Experiments

4.1. Implementation details

In this work, we select our previous work ViTAStereo [36] as the student network and formulate our self-supervised framework as Self-DispArbiter. ViTAStereo is able to perform stereo matching using a frozen pretrained Depth Anything 3 [32] model. Therefore, relative depth maps are accessed efficiently by deploying a separate dense prediction Transformer [37] head. During the self-supervised training process of ViTAStereo, the weights of both the DDCN and the Depth Anything 3 model are frozen, with only the disparity decoder and a lightweight feature adapter within ViTAStereo trainable. During disparity inference, relative depth and disparity confidence estimation are not required.

Self-DispArbiter and DDCN are pre-trained on the synthetic SceneFlow dataset [38] and evaluated on the real-world KITTI Stereo 2012 [39], 2015 [40], and Middlebury [41] datasets. Unless otherwise specified, experiments are conducted on the Middlebury training set because of its high-quality disparity ground truth and complex environments. To demonstrate the ability of our sparsified D3Stereo to generate reliable teacher disparities, we generate low-quality matching cost volumes as input using MCCNN [42]. All experiments are conducted on an NVIDIA RTX 4090 GPU.

4.2. Hyperparameter selection in DispArbiter

We first investigate the selection of the conflicting disparity threshold τ and the ratio hyperparameter γ in the CAF by alternatively employing IGEVStereo and CreStereo as the teacher and student networks. Fig. 3 illustrates the end-point error (EPE, denoted by ε, in px) improvements of the enhanced disparities $\boldsymbol{D}^{\text{E}}$ over the student disparities $\boldsymbol{D}^{\text{S}}$. These improvements are evaluated at pseudo label locations ($\triangle\varepsilon_P = \varepsilon_P^{\text{S}} - \varepsilon_P^{\text{E}}$) and across all pixels ($\triangle\varepsilon = \varepsilon^{\text{S}} - \varepsilon^{\text{E}}$). We also report the proportion of retained pseudo labels, denoted by κ_{P} (in %).

As observed, all configurations yield a positive $\triangle\varepsilon_P$, indicating superior accuracy of pseudo labels over the corresponding student disparities. Furthermore, increasing both τ and γ leads to larger improvements in pseudo label accuracy $\triangle\varepsilon_P$ but lower pseudo label densities κ_{P}. The maximum overall accuracy improvement $\triangle\varepsilon$ occurs at $\tau = 0.5$ and $\gamma = 0$. To strike a balance between pseudo label accuracy and density, we set $\tau = 3$ and $\gamma = 1$ without manual tuning for all subsequent experiments.

4.3. Ablation study and cross validation

We first conduct ablation studies on our local pattern volume alignment function $\mathcal{A}(\cdot)$ in (1) and activation function $\mathcal{F}(\cdot)$ in (2), and compare $\mathcal{L}_{\text{LER}}$ with the global distribution loss $\mathcal{L}_{\text{GD}}$ used in SEDNet [13]. We generate the density-EPE receiver operating characteristic (ROC) curves and then use the area under the curve (AUC) [25] to evaluate the disparity confidence estimation accuracy. The quantitative results in Table 1 show that our $\mathcal{L}_{\text{LER}}$ reduces the AUC by 3.4% to 5.3% compared with $\mathcal{L}_{\text{GD}}$. Removing $\mathcal{A}(\cdot)$ and $\mathcal{F}(\cdot)$ leads to an increase in AUC by 3.3% to 7.4% and 13.3% to 22.1%, respectively. By mitigating excessive variations at depth discontinuities, our $\mathcal{F}(\cdot)$ helps preserve local patterns in the subsequent normalization process, thereby enabling more accurate disparity confidence estimation.

We also evaluate our proposed SES through cross-validation experiments using IGEVStereo [44], CreStereo [43], and our Self-DispArbiter. The quantitative results in Table 2 show that DispArbiter exhibits robust pseudo label mining capability, particularly when provided with comparable or slightly poorer teacher disparities. This is further supported by the qualitative results in Fig. 4. These findings also underscore the significance of our sparsified D3Stereo in generating sparse yet highly accurate teacher disparities from low-quality matching costs. Furthermore, SES leads to a favorable and stable trade-off, discarding an acceptable proportion of pseudo labels (less than 7.2%) to secure a significant improvement in accuracy (up to 9.8%). This accuracy improvement is particularly pronounced when mining pseudo labels from the less accurate teacher disparities generated by Self-DispArbiter.

Finally, we evaluate different ViT-Large Depth Anything models for relative depth generation in DDCN. The quantitative results presented in Table 4 show that relative depth maps from Depth Anything 3 [32] lead to more accurate disparity confidence than those from Depth Anything V2 [33]. Different Depth Anything 3 variants achieve comparable performance. Consequently, we adopt the relative depth maps generated by the MONO-LARGE Depth Anything 3 model in this work.

4.4. Comparison with SoTA

4.4.1. Conflict Arbitration between supervised methods

To explore the performance upper bound of DispArbiter, we mine pseudo labels for the recent SoTA method, Monster++ [46], using its

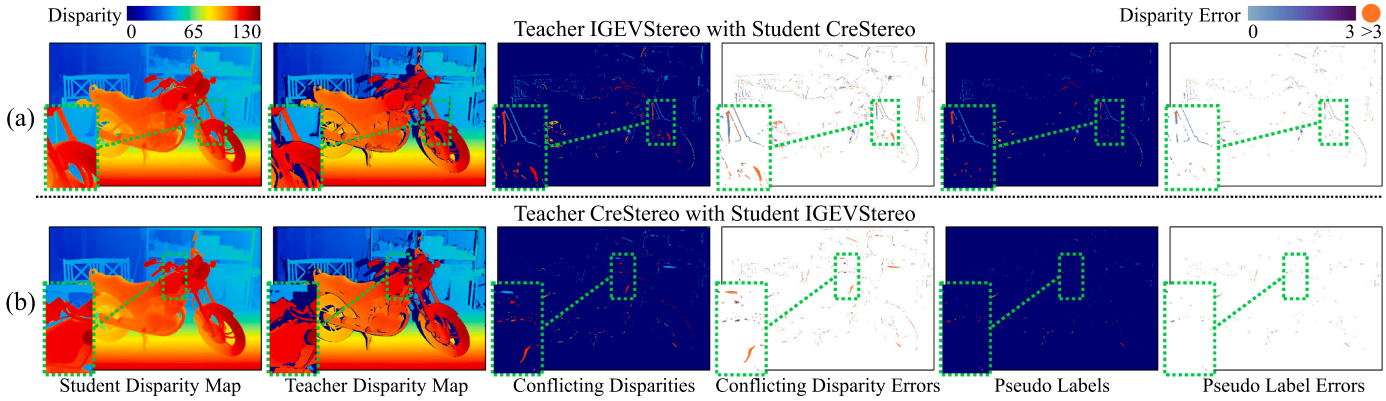

Fig. 4. Cross-validation results for Self-DispArbiter, CreStereo [43], and IGEVStereo [44].

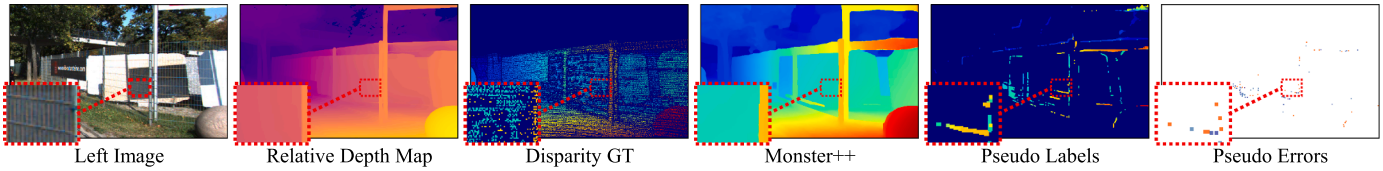

Fig. 5. Failures of DispArbiter while mining pseudo disparity labels from Monster [45] to Monster++ [46] in mesh regions on the KITTI Stereo 2015 benchmark.

Table 2
Ablation studies and cross-validation results for the sparse embedding strategy.

Teacher \ Student	Self-DispArbiter ($\epsilon = 1.81$)				CreStereo ($\epsilon = 0.83$)				IGEVStereo ($\epsilon = 0.74$)			
	w/o SES		w/ SES		w/o SES		w/ SES		w/o SES		w/ SES	
	$\triangle\epsilon_P$	κ_P	$\triangle\epsilon_P$	κ_P	$\triangle\epsilon_P$	κ_P	$\triangle\epsilon_P$	κ_P	$\triangle\epsilon_P$	κ_P	$\triangle\epsilon_P$	κ_P
Self-DispArbiter ($\epsilon = 1.81$)	–	–	–	–	−0.08	0.20	0.11	0.19	−3.38	0.17	−3.05	0.16
CreStereo [43] ($\epsilon = 0.83$)	11.5	1.30	11.6	1.24	–	–	–	–	3.44	0.05	3.45	0.05
IGEVStereo [44] ($\epsilon = 0.74$)	12.5	1.38	12.61	1.28	7.94	0.16	8.04	0.15	–	–	–	–

Table 3
Pseudo disparity label mining from Monster [45] to Monster++ [46].

Method	Middlebury				KITTI Stereo 2012				KITTI Stereo 2015					
	ALL Pixels		Noc Pixels		All Pixels		Noc Pixels		All Pixels			Noc Pixels		
	Bad2.0	Bad4.0	Bad2.0	Bad4.0	Out-2	Out-3	Out-2	Out-3	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all
Monster [45]	**6.14**	–	2.64	–	1.75	1.09	1.36	0.84	1.13	–	1.41	1.05	–	1.33
Monster++ [46]	6.40	3.52	2.60	1.18	1.70	1.07	1.30	0.79	1.12	**2.65**	1.37	1.02	**2.67**	1.29
DispArbiter (Ours)	6.38	**3.50**	**2.57**	**1.17**	**1.69**	**1.06**	**1.29**	**0.78**	**1.10**	2.69	**1.37**	**1.00**	2.70	**1.28**

Table 4
Ablation studies on Depth Anything models for generating relative depth maps in DDCN.

Dataset	Depth Anything 3				Depth Anything V2	Optimal
	MONO-LARGE	METRIC-LARGE	LARGE	LARGE-1.1	LARGE	
Middlebury	0.90	**0.89**	0.91	0.92	0.92	0.30
KITTI	**0.49**	0.50	**0.49**	0.50	0.52	0.21

Table 5
Comparisons with SoTA self-supervised stereo matching methods on the Middlebury dataset in terms of EPE. The best results are shown in bold type.

Method	Adi.	Art.	Jad.	Mot.	Mot.E	Pia.	Pia.L	Pip.P	Playr.	Playt	Playt.P	Rec.	She.	Ted.	Vin.	**Avg.**
TCSCSM [47]	2.33	3.06	11.9	2.68	2.35	3.56	6.49	5.34	3.20	19.5	2.35	1.91	9.29	1.42	7.90	4.81
UnDAF-GANet [30]	**0.55**	2.97	**4.98**	1.65	1.57	2.23	**3.01**	2.15	2.02	1.90	1.71	1.23	2.89	0.93	**1.63**	2.06
DualNet [8]	1.26	3.06	29.6	2.09	2.19	2.45	6.28	3.66	3.12	**1.69**	1.40	1.19	5.61	1.07	7.03	4.79
UnViTAStereo [25]	1.23	3.41	56.4	2.75	2.57	1.53	3.55	4.99	2.46	2.81	1.41	1.40	4.34	1.02	16.6	7.32
Self-DispArbiter	0.70	**1.36**	5.40	**1.54**	**1.30**	**1.14**	3.73	**1.85**	**1.83**	3.58	**0.86**	**0.84**	**2.55**	**0.72**	4.28	**1.90**

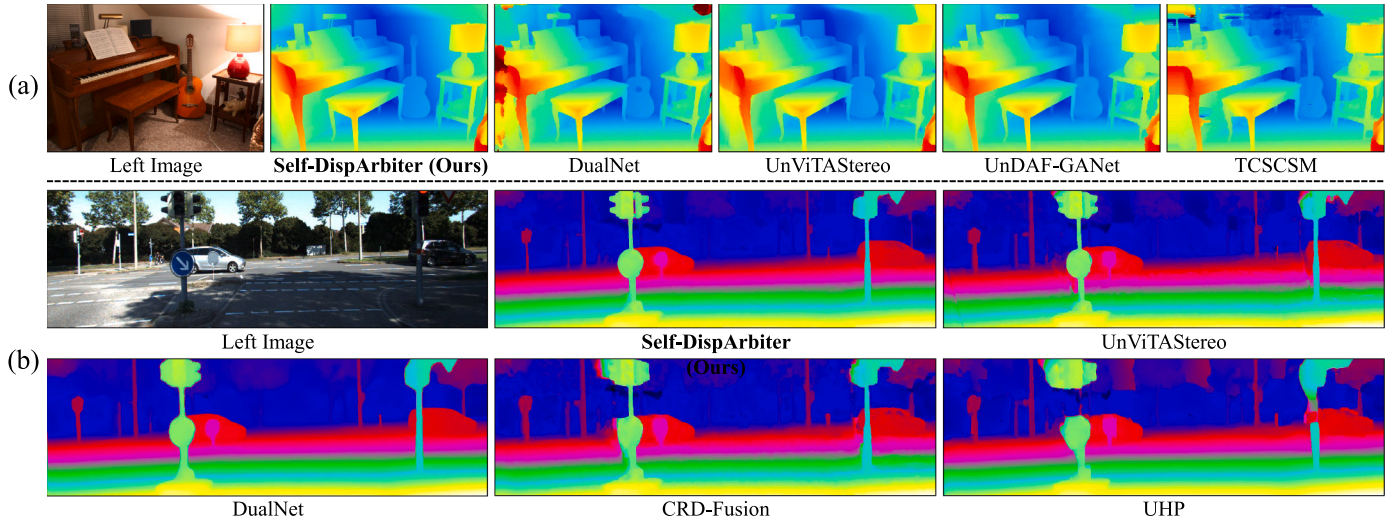

Fig. 6. Comparisons of Self-DispArbiter with SoTA self-supervised methods on the (a) Middlebury and (b) KITTI stereo benchmarks.

Table 6
Comparisons with SoTA self-supervised stereo matching methods on the KITTI stereo benchmarks.

Network	KITTI Stereo 2012				KITTI Stereo 2015				Runtime
	All Pixels		Noc Pixels		All Pixels		Noc Pixels		
	EPE	Out-3	EPE	Out-3	D1-bg	D1-all	D1-bg	D1-all	
Flow2Stereo [48]	1.1	5.11	1.0	4.58	5.01	6.61	4.77	6.29	**0.05**
PASMnet_192 [26]	–	–	–	–	5.41	7.23	5.02	6.69	0.50
OASM-DDS [49]	1.2	5.69	1.1	4.81	4.46	6.51	–	6.02	–
Tong et al. [50]	–	–	–	–	6.55	8.11	5.86	7.32	0.17
Permutation-Stereo [51]	1.8	8.48	1.6	7.39	5.53	7.18	5.18	6.72	30.0
CRD-Fusion [10]	1.1	5.40	0.9	4.38	4.59	6.11	4.30	5.69	**0.02**
UHP [52]	1.3	7.09	1.2	6.05	5.00	6.45	4.65	5.93	**0.02**
SPSMnet [53]	–	–	–	–	5.42	6.65	4.94	6.10	0.80
DualNet [8]	**0.6**	**2.59**	**0.6**	**2.06**	**2.46**	**2.92**	**2.28**	**2.67**	0.17
UnViTAStereo [25]	0.9	4.16	0.8	3.46	3.58	5.03	3.28	4.61	0.22
Self-DispArbiter (Ours)	0.7	2.94	**0.6**	2.39	2.77	3.80	2.54	3.45	0.22

weaker version, Monster [45]. As shown in Table 3, we first mine pseudo disparity labels from the low-quality disparities generated by Monster and then substitute these labels into the high-quality disparity map generated by Monster++. The output enhanced disparity maps $\boldsymbol{D}^{\mathrm{E}}$ outperform those from Monster++ on both the Middlebury and KITTI Stereo benchmarks, yielding accuracy improvements of 0.3% to 1.2% on Middlebury and 0.6% to 2.0% on KITTI Stereo 2012. These results demonstrate the potential of DispArbiter to generate pseudo disparity labels for unlabeled stereo images. The slight performance drop on the KITTI 2015 D1-fg metric likely stems from incorrect relative depth estimates, as visualized in Fig. 5. Specifically, inaccurate relative depth estimates in mesh regions mislead the DDCN, triggering a cascade of flawed confidence evaluations that inevitably result in the retention of erroneous teacher disparities as pseudo labels. Remarkably, DispArbiter ranks 1st on the KITTI Stereo benchmarks.

4.4.2. *Self-DispArbiter*

Quantitative and qualitative experimental comparisons between Self-DispArbiter and other self-supervised methods are presented in Tables 5, 6, and Fig. 6, respectively. On the Middlebury benchmark, Self-DispArbiter achieves the highest accuracy, reducing the EPE by 16% to 80% compared with existing methods. On the KITTI benchmark, our method surpasses all other self-supervised approaches except DualNet, trailing by 0% to 14% on KITTI 2012 and 11% to 23% on KITTI 2015. We attribute the accuracy inconsistency between the Middlebury and KITTI benchmarks to DualNet's additional modeling of the probability distribution of teacher disparities and the erroneous relative depth estimates discussed in Section 4.4.1. In general, the remarkable disparity accuracy achieved by Self-DispArbiter highlights the potential importance of improved pseudo-label generation methods, particularly in the field of self-supervised stereo matching.

Table 7
Comparisons among disparity confidence estimation methods.

Method	Dataset			
	Middlebury		KITTI	
	AUC	Runtime (s)	AUC	Runtime (s)
OTB [16]	1.28	2.07	0.71	1.12
DPS [24]	0.97	7.14	0.50	4.71
DDCV [25]	0.96	0.35	0.52	0.19
DDCN (ours)	0.86	0.21	0.49	0.12
Optimal	0.30	–	0.21	–

4.4.3. *Disparity confidence estimation*

We compare our proposed DDCN against recent SoTA disparity confidence estimation methods compatible with black-box stereo matching. The input disparity maps are generated using our Self-DispArbiter. Quantitative and qualitative results are presented in Table 7 and Fig. 7, respectively. It can be observed that DDCN reduces the AUC by 29.7% to 31.0% compared to OTB [16], 2.0% to 11.3% compared to DPS [24],

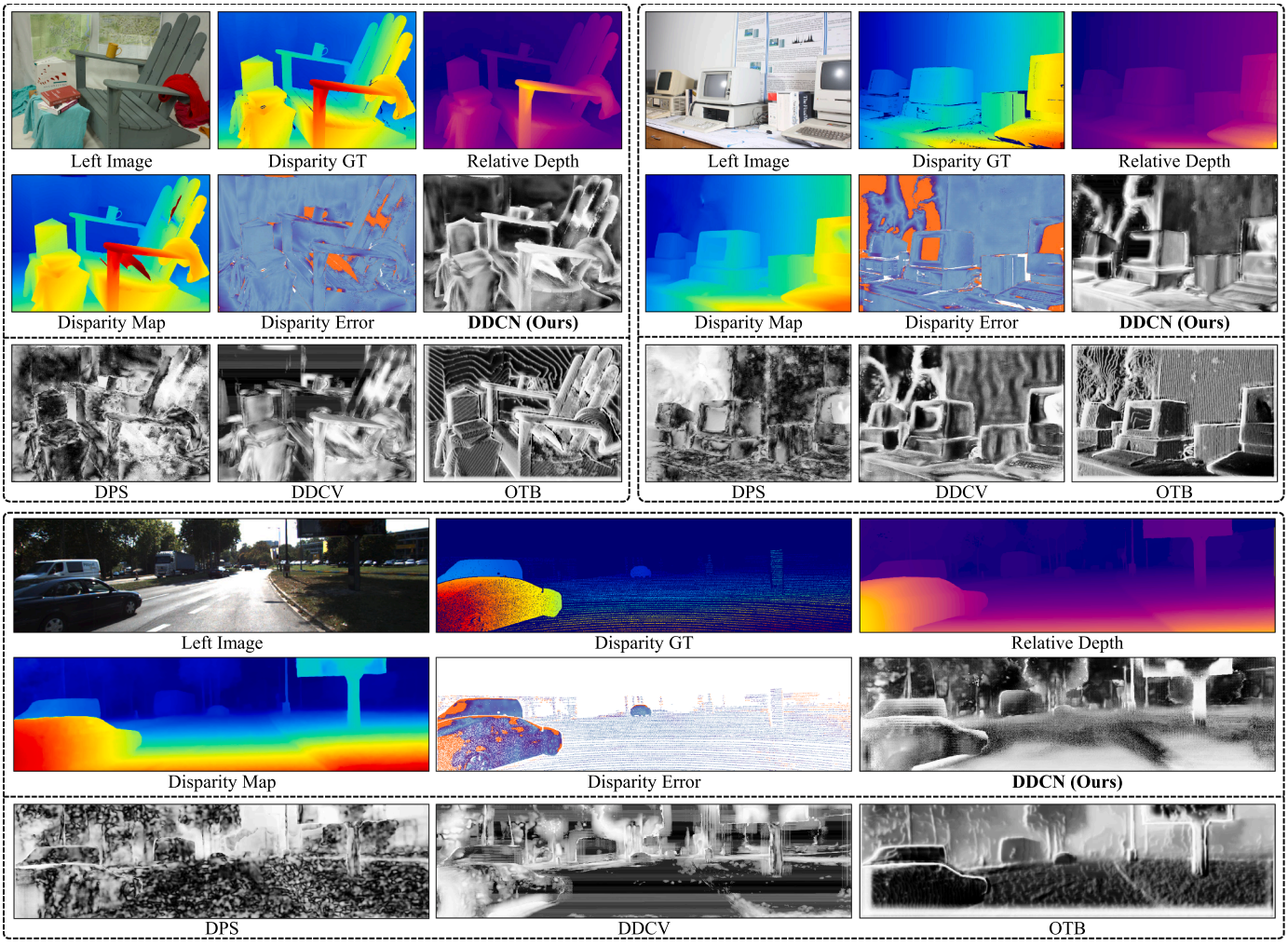

Fig. 7. Comparisons between disparity confidence estimation methods that are compatible with black-box stereo matching.

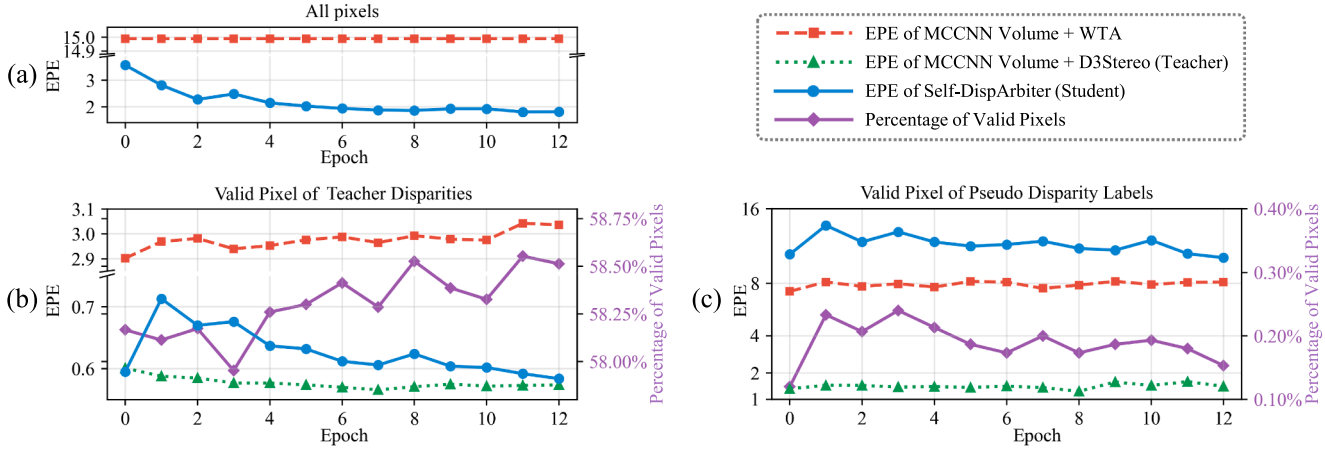

Fig. 8. Quantitative results of the self-supervised learning process of Self-DispArbiter.

and 5.8% to 6.3% compared to our previous work DDCV [25]. It is noteworthy that DPS [24] replaces the intermediate stereo matching data with a disparity volume, which is generated by performing a plane sweep on the right image based on shifted disparity prediction results, followed by multiple stereo matching operations (four additional disparity inferences in our experiments, following the default configuration in the official code). These multiple stereo matching inference passes make DPS highly inefficient and less practical.

However, qualitative results in Fig. 7 expose the limitations of DDCN in large regions with extensive disparity errors, where local disparity patterns lack indicative disparity correspondences that are consistent with the 3D geometric priors provided by the relative depth map. Addi-

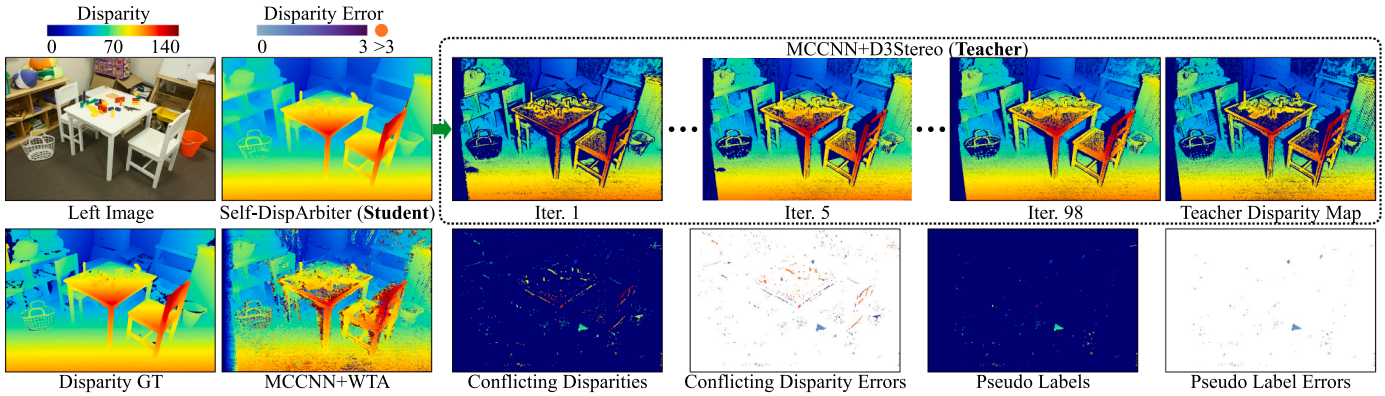

Fig. 9. Qualitative results of the self-supervised learning process of Self-DispArbiter.

tionally, DDCN predicts lower confidence in regions with considerable depth discontinuities. We attribute this to the scale inconsistency between the disparity map and the relative depth map.

4.5. Discussion

We conduct a detailed analysis of the intermediate results in the self-supervised training process of Self-DispArbiter, with quantitative results and visual illustrations provided in Figs. 8 and 9, respectively. Results obtained by applying the Winner-Take-All (WTA) strategy to MCCNN cost volumes are also included to demonstrate the effectiveness of D3Stereo. As shown in Fig. 8(a), applying MCCNN with WTA yields noisy disparities with low accuracy, whereas the stereo matching accuracy of Self-DispArbiter progressively improves and ultimately converges during training. Fig. 8(b) evaluates performance on the valid pixels of teacher disparities generated by D3Stereo. Notably, D3Stereo produces teacher disparities with even higher accuracy than Self-DispArbiter across more than half of all pixels. Furthermore, the accuracy of MCCNN with WTA also improves significantly at these valid pixels, demonstrating the capability of D3Stereo to identify low-noise regions from cost volumes. As self-supervised training progresses, the student network improves in accuracy, which in turn provides the teacher with a larger number of more precise initial decisive disparities. Consequently, both the quantity and the accuracy of the output teacher disparities increase, as reflected in Fig. 8(b) by the rising number of valid pixels and the decreasing EPE of the teacher disparities. Fig. 8(c) presents statistics evaluated at the pseudo-label locations, where the disparity error of Self-DispArbiter is approximately 30% higher than that of MCCNN with WTA. This observation further validates the student deficiency awareness of DispArbiter. As the student network continues to improve over the course of self-supervised training, DispArbiter gradually becomes less able to identify additional pseudo disparity labels that provide useful supervision for the student. This corresponds to the declining trend of pseudo-label counts observed in Fig. 8(c).

Moreover, Fig. 8(c) also reveals that despite the accuracy convergence of Self-DispArbiter, a substantial gap remains between the accuracy of the pseudo labels and that of the student disparities. This indicates that existing training strategies or student network architectures fail to fully exploit the potential of pseudo labels generated by DispArbiter. We attribute this to knowledge forgetting in the student network in regions lacking pseudo labels. In general, we argue that the critical bottleneck of self-supervised methods lies in effectively assimilating sparse yet accurate pseudo labels while preventing representation collapse in unlabeled regions.

5. Conclusion

In this work, we propose DispArbiter, a novel framework designed to mine pseudo disparity labels from teacher disparities with student deficiency awareness, alongside a sparsified D3Stereo method to generate sparse yet highly accurate teacher disparities. By resolving disparity conflicts among SoTA stereo matching methods, DispArbiter mines accurate pseudo disparities from low-quality teacher disparities to further improve student predictions that are already more accurate than those of the teacher. The collaboration between DispArbiter and D3Stereo achieves SoTA accuracy in self-supervised stereo matching. This research underscores the critical importance of (1) more accurate disparity confidence estimation methods compatible with black-box stereo matching and (2) approaches designed for performing sparse yet reliable stereo matching. Furthermore, integrating future advancements in relative depth and disparity confidence estimation would further enhance the performance of DispArbiter. In future work, we aim to establish a multi-expert discriminative system based on DispArbiter to generate pseudo ground-truth disparities for unlabeled stereo images.

CRediT authorship contribution statement

Chuang-Wei Liu: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Mengtan Zhang:** Visualization, Validation, Methodology; **Nachuan Ma:** Methodology, Investigation; **Jin Wu:** Methodology, Investigation; **Hanli Wang:** Writing – review & editing, Resources, Project administration; **Qijun Chen:** Methodology; **Rui Fan:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the National Natural Science Foundation of China under Grant 62473288 and Grant 62233013, the Explorers Program of Shanghai (Basic Research Funding) under Grant 26TS1404200, the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University (No. HMHAI-202406), the Fundamental Research Funds for the Central Universities, NIO University Programme (NIO UP), and the Xiaomi Young Talents Program.

Data availability

Data will be made available on request.

References

[1] Z. Wu, et al., S^3M-Net: joint learning of semantic segmentation and stereo matching for autonomous driving, IEEE Trans. Intell. Veh. 9 (2) (2024) 3940–3951.

[2] S. Guo, et al., LIX: implicitly infusing spatial geometric prior knowledge into visual semantic segmentation for autonomous driving, IEEE Trans. Image Process. 34 (2025) 7250–7263.

[3] G. Tang, et al., TiCoSS: tightening the coupling between semantic segmentation and stereo matching within a joint learning framework, IEEE Trans. Autom. Sci. Eng. 22 (2025) 18646–18658.

[4] C.-W. Liu, et al., Stereo matching: fundamentals, state-of-the-art, and existing challenges, in: Autonomous Driving Perception: Fundamentals and Applications, Springer, 2023, pp. 63–100.

[5] Y. Guo, et al., Learning deformable hypothesis sampling for patchmatch multi-view stereo in the wild, Inf. Fusion 113 (2025) 102646.

[6] R. Fan, et al., Road surface 3D reconstruction based on dense subpixel disparity map estimation, IEEE Trans. Image Process. 27 (6) (2018) 3025–3035.

[7] R. Fan, et al., Rethinking road surface 3-D reconstruction and pothole detection: from perspective transformation to disparity map segmentation, IEEE Trans. Cybern. 52 (7) (2021) 5799–5808.

[8] Y. Wang, et al., DualNet: robust self-supervised stereo matching with pseudo-label supervision, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2025, pp. 8178–8186.

[9] P. Xu, et al., Unsupervised stereo via multi-baseline geometry-consistent self-training, Comput. Res. Repos. (CoRR) arXiv:2508.10838v2 (2026).

[10] X. Fan, et al., Occlusion-aware self-supervised stereo matching with confidence guided raw disparity fusion, in: 2022 19th Conference on Robots and Vision (CRV), IEEE, 2022, pp. 132–139.

[11] H. Wang, et al., PVStereo: pyramid voting module for end-to-end self-supervised stereo matching, IEEE Robot. Autom. Lett. 6 (3) (2021) 4353–4360. https://doi.org/10.1109/LRA.2021.3068108

[12] Z. Shen, et al., Digging into uncertainty-based pseudo-label for robust stereo matching, IEEE Trans. Pattern Anal. Mach. Intell. 45 (12) (2023) 14301–14320.

[13] L. Chen, et al., Learning the distribution of errors in stereo matching for joint disparity and uncertainty estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17235–17244.

[14] J. Zhou, et al., Consistency-aware self-training for iterative-based stereo matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 16641–16650.

[15] M. Poggi, et al., Self-adapting confidence estimation for stereo, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2020, pp. 715–733.

[16] M. Poggi, et al., Learning a confidence measure in the disparity domain from O(1) features, Comput. Vis. Image Underst. 193 (2020) 2905–2913.

[17] X. Hu, P. Mordohai, A quantitative evaluation of confidence measures for stereo vision, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2121–2133.

[18] M. Poggi, et al., Quantitative evaluation of confidence measures in a machine learning world, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5228–5237.

[19] M. Poggi, S. Mattoccia, Learning from scratch a confidence measure, in: Proceedings of 27th British Machine Vision Conference (BMVC), 2016, pp. 1–13.

[20] F. Tosi, et al., Beyond local reasoning for stereo confidence estimation with deep learning, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 319–334.

[21] O. Ronneberger, et al., U-Net: convolutional networks for biomedical image segmentation, in: Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, 2015, pp. 234–241.

[22] M. Mehltretter, Joint estimation of depth and its uncertainty from stereo images using Bayesian deep learning, ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 2 (2022) 69–78.

[23] M. Mehltretter, C. Heipke, Aleatoric uncertainty estimation for dense stereo matching via CNN-based cost volume analysis, ISPRS J. Photogramm. Remote Sens. 171 (2021) 63–75.

[24] J.Y. Lee, et al., Modeling stereo-confidence out of the end-to-end stereo-matching network via disparity plane sweep, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2024, pp. 2901–2910.

[25] C.-W. Liu, et al., Integrating disparity confidence estimation into relative depth prior-guided unsupervised stereo matching, IEEE Trans. Circuits Syst. Video Technol. 36 (1) (2026) 350–362.

[26] L. Wang, et al., Parallax attention for unsupervised stereo correspondence learning, IEEE Trans. Pattern Anal. Mach. Intell. 44 (4) (2020) 2108–2125.

[27] R. Haeusler, R. Nair, D. Kondermann, Ensemble learning for confidence measures in stereo vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 305–312.

[28] M.-G. Park, K.-J. Yoon, Learning and selecting confidence measures for robust stereo matching, IEEE Trans. Pattern Anal. Mach. Intell. 41 (6) (2018) 1397–1411.

[29] X. Yue, et al., Semi-stereo: a universal stereo matching framework for imperfect data via semi-supervised learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 646–655.

[30] H. Wang, et al., UnDAF: a general unsupervised domain adaptation framework for disparity or optical flow estimation, in: International Conference on Robotics and Automation (ICRA), IEEE, 2022, pp. 01–07.

[31] P. Xu, et al., Self-supervised stereo matching with multi-baseline contrastive learning, Comput. Res. Repos. arXiv:2508.10838 (2025).

[32] H. Lin, et al., Depth Anything 3: recovering the visual space from any views, Comput. Res. Repos. arXiv:2511.10647 (2025).

[33] L. Yang, et al., Depth Anything V2, Adv. Neural Inf. Process. Syst. 37 (2024) 21875–21911.

[34] T. Ojala, et al., Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.

[35] X. Guo, et al., Group-wise correlation stereo network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3273–3282.

[36] C.-W. Liu, et al., Playing to vision foundation model's strengths in stereo matching, IEEE Trans. Intell. Veh. (2024). https://doi.org/10.1109/TIV.2024.3467287.

[37] R. Ranftl, A. Bochkovskiy, V. Koltun, Vision transformers for dense prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12179–12188.

[38] N. Mayer, et al., A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4040–4048.

[39] A. Geiger, et al., Are we ready for autonomous driving? The KITTI vision benchmark suite, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 3354–3361.

[40] M. Menze, A. Geiger, Object scene flow for autonomous vehicles, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3061–3070.

[41] D. Scharstein, et al., High-resolution stereo datasets with subpixel-accurate ground truth, in: Pattern Recognition: 36th German Conference (GCPR), Springer, 2014, pp. 31–42.

[42] J. Zbontar, Y. LeCun, Computing the stereo matching cost with a convolutional neural network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1592–1599.

[43] J. Li, et al., Practical stereo matching via cascaded recurrent network with adaptive correlation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16263–16272.

[44] G. Xu, et al., Iterative geometry encoding volume for stereo matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21919–21928.

[45] J. Cheng, et al., Monster: marry monodepth to stereo unleashes power, in: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 6273–6282.

[46] J. Cheng, et al., MonSter++: unified stereo matching, multi-view stereo, and real-time stereo with monodepth priors, Comput. Res. Repos. arXiv:2501.08643 (2025).

[47] C. Cheng, et al., An unsupervised stereo matching cost based on sparse representation, Int. J. Wavelets Multiresolution Inf. Process. 19 (01) (2021) 2050060.

[48] P. Liu, et al., Flow2Stereo: effective self-supervised learning of optical flow and stereo matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6648–6657.

[49] A. Li, et al., Unsupervised occlusion-aware stereo matching with directed disparity smoothing, IEEE Trans. Intell. Transp. Syst. 23 (7) (2021) 7457–7468.

[50] K.W. Tong, et al., Adaptive cost volume representation for unsupervised high-resolution stereo matching, IEEE Trans. Intell. Veh. 8 (1) (2022) 912–922.

[51] P.-A. Brousseau, S. Roy, A permutation model for the self-supervised stereo matching problem, in: 2022 19th Conference on Robots and Vision (CRV), IEEE, 2022, pp. 122–131.

[52] R. Yang, et al., Unsupervised hierarchical iterative tile refinement network with 3D planar segmentation loss, IEEE Robot. Autom. Lett. 9 (3) (2024) 2678–2685.

[53] X. Yang, et al., Self-supervised learning of PSMNet via generative adversarial networks, in: International Conference on Intelligent Computing (ICIC), Springer, 2024, pp. 469–479.